D1064–D1070 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 27 October 2014*
doi: 10.1093/nar/gku1002

HAMAP in 2015: updates to the protein family classification and annotation system

Ivo Pedruzzi[1,†], Catherine Rivoire[1,†], Andrea H. Auchincloss[1], Elisabeth Coudert[1], Guillaume Keller[1], Edouard de Castro[1], Delphine Baratin[1], Béatrice A. Cuche[1], Lydie Bougueleret[1], Sylvain Poux[1], Nicole Redaschi[1], Ioannis Xenarios[1,2,3,4] and Alan Bridge[1,*]

[1]Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland, [2]Vital-IT Group, SIB Swiss Institute of Bioinformatics, CH-1015, Lausanne, Switzerland, [3]Center for Integrative Genomics, University of Lausanne, CH-1015, Lausanne, Switzerland and [4]Department of Biochemistry, University of Geneva, CH-1211 Geneva 4, Switzerland

Received September 07, 2014; Revised October 6, 2014; Accepted October 07, 2014

ABSTRACT

HAMAP (High-quality Automated and Manual Annotation of Proteins—available at http://hamap.expasy.org/) is a system for the automatic classification and annotation of protein sequences. HAMAP provides annotation of the same quality and detail as UniProtKB/Swiss-Prot, using manually curated profiles for protein sequence family classification and expert curated rules for functional annotation of family members. HAMAP data and tools are made available through our website and as part of the UniRule pipeline of UniProt, providing annotation for millions of unreviewed sequences of UniProtKB/TrEMBL. Here we report on the growth of HAMAP and updates to the HAMAP system since our last report in the NAR Database Issue of 2013. We continue to augment HAMAP with new family profiles and annotation rules as new protein families are characterized and annotated in UniProtKB/Swiss-Prot; the latest version of HAMAP (as of 3 September 2014) contains 1983 family classification profiles and 1998 annotation rules (up from 1780 and 1720). We demonstrate how the complex logic of HAMAP rules allows for precise annotation of individual functional variants within large homologous protein families. We also describe improvements to our web-based tool HAMAP-Scan which simplify the classification and annotation of sequences, and the incorporation of an improved sequence-profile search algorithm.

INTRODUCTION

Falling costs and continuing technological advances in DNA sequencing have led to an explosion in the number of available whole genome sequences from all branches of the tree of life, opening up exciting new possibilities for research into the evolution and function of biological systems. However as the number of protein-coding gene sequences continues to grow exponentially, the tiny fraction of experimentally characterized sequences continues to shrink—this despite the best efforts of groups such as the Enzyme Function Initiative (1) and COMBREX (2) to accelerate the rate of functional characterization through combined computational and experimental approaches. This growing gap highlights a need for automated systems that can effectively leverage the available experimental information to provide precise functional annotation for the tens of millions of predicted protein sequences that will probably never be characterized (3).

One such system is HAMAP (High-quality Automated and Manual Annotation of Proteins), which provides automatic classification and functional annotation of protein sequences based on their homology to characterized templates (4). HAMAP is based on a collection of expert curated protein family profiles, which are used to determine family membership of protein sequences, and annotation rules, which specify the appropriate annotation for family members. HAMAP rules permit the annotation of protein sequences to the same level of detail and quality as manually curated UniProtKB/Swiss-Prot records, annotating protein and gene names, function, catalytic activity, cofactors, subcellular location, protein–protein interactions, as well as sequence features such as the presence of specific domains, motifs and functionally important sites (such as ion-, substrate- and cofactor-binding sites, catalytic residues

*To whom correspondence should be addressed. Tel: +41 22 379 5059; Fax: +41 22 379 5858; Email: alan.bridge@isb-sib.ch
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

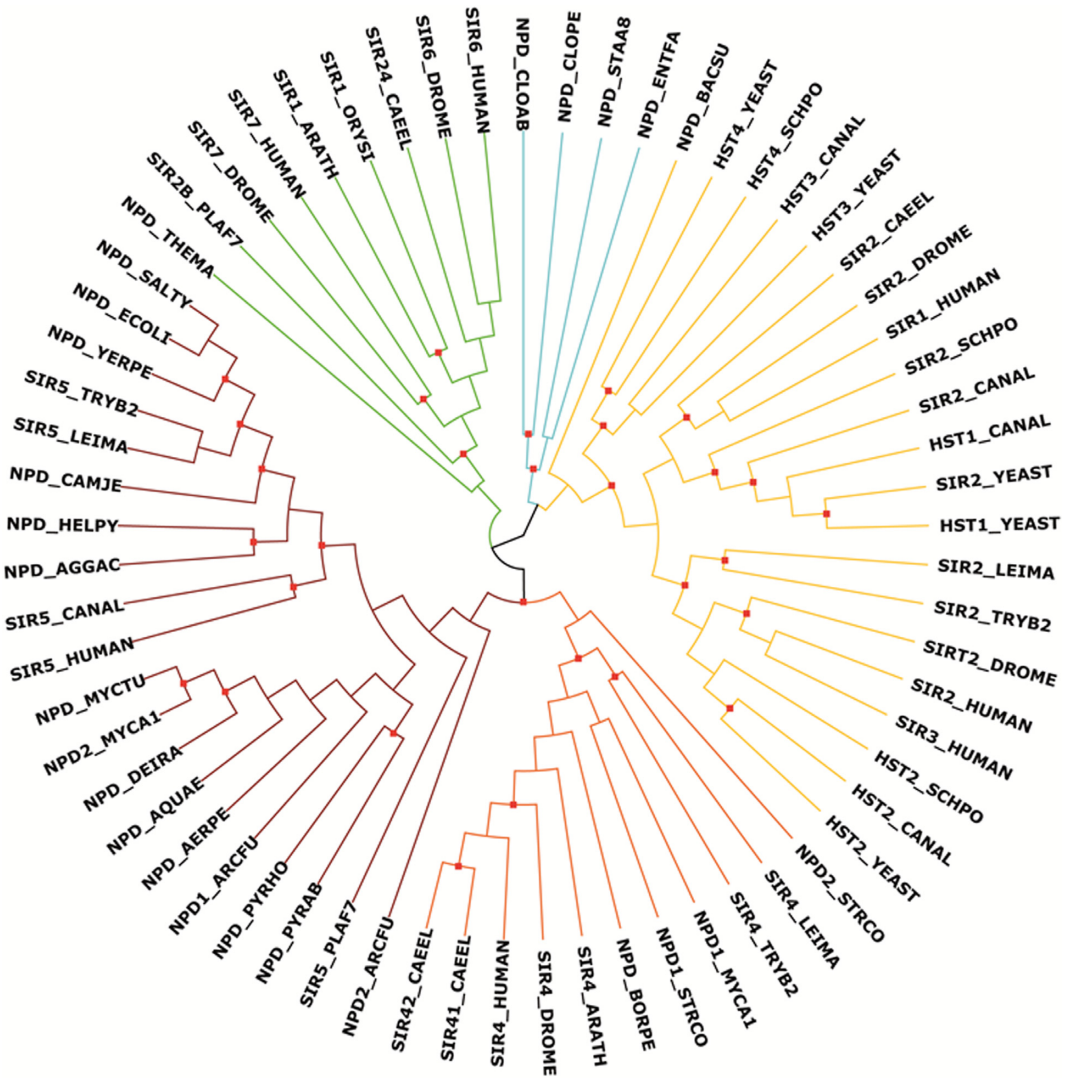

Figure 1. Maximum likelihood cladogram of the sirtuin superfamily. Maximum likelihood (ML) analyses of selected sirtuin family members resulted in 12 trees with two distinct topologies for the main classes I-IV and U, suggesting either classes II and III or classes II and VI to be sister clades. The tree topology with highest branch support is shown. Branches are colored according to families: class I = dark yellow, class II = orange, class III = red, class IV = green, class U = cyan. Branches with aLRT SH-like support values of 0.9 or higher are marked by a red dot**.** Methods**:** 65 sirtuin protein family members from 33 species were aligned with MAFFT (21) (version 7; parameters: L-INS-i, JTT200). From the alignment, we selected manually homologous regions using the alignment editor Jalview (22); three data models were created with a length of 238, 220 and 193 amino acids, respectively. The best fitting model of protein evolution was determined with ProtTest (23) (version 3.2; parameters: fixed BIONJ tree calculated under the JTT model of amino acid substitution; rate variation; amino acid frequencies to be the LG model plus gamma distribution). Maximum likelihood (ML) phylogenies and ML consensus trees from 100 bootstrap replicates were inferred with PhyML (24) (version 3.0) and RAxML (25) (version 7.2.8). The tree was visualized with Archaeopteryx (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx). Protein sequences and multiple sequence alignments are provided in supplementary file S2.

and post-translational modifications). Annotations are provided in the form of the human-readable UniProtKB text format and using UniProt controlled vocabularies and terms from the Gene Ontology (GO) (5). As well as the annotations themselves, HAMAP rules also specify the conditions under which these annotations may be applied, such as a requirement for key functional residues (identified by structural or other experimental studies). Such conditions can reduce the incidence of erroneous annotation, particularly in large, functionally diverse families—errors that tend to persist in public sequence databases (6–8).

HAMAP forms one component of the UniProt UniRule system that provides annotation for the unreviewed component of the UniProt Knowledgebase UniProtKB/TrEMBL (9). HAMAP family profiles and annotation rules are created (and updated) concurrently with the curation of experimentally characterized templates into UniProtKB/Swiss-Prot, by the same expert curators. This ensures that the family profiles accurately reflect the properties of trusted protein family members, that target sequences are annotated to the quality standards of UniProtKB/Swiss-Prot, and that updates to UniProtKB/Swiss-Prot records are subsequently recorded in HAMAP rules (and propagated to

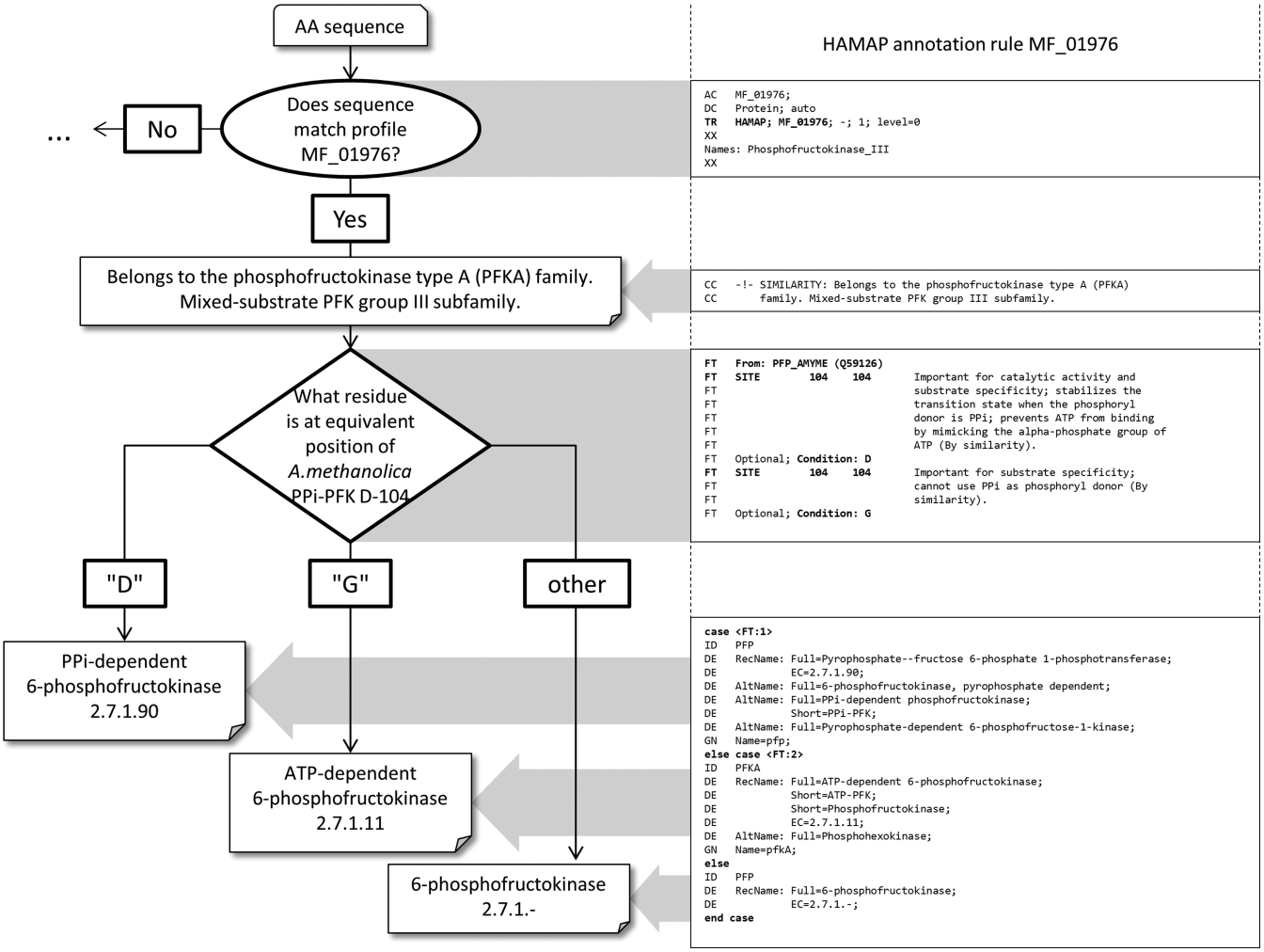

Figure 2. HAMAP annotation rule MF_01976 for mixed-substrate PFK group III family. The right hand panel shows snippets of the annotation rule MF_01976 including conditions used to specify site-specific annotations propagated to target sequences. If a protein sequence matches the HAMAP family profile MF_01976, then appropriate annotations for all members of that family (such as family membership) are attached to the sequence. For the annotation of sequence features, the target sequence is aligned to the seed alignment and the active site residue from the template sequence mapped to the target sequence. The nature of the residue at the equivalent position in the target sequence determines which of the possible conditional annotations will be attached to the sequence.

homologous UniProtKB/TrEMBL records). In addition to UniProtKB, HAMAP also provides protein family annotation for Ensembl Genomes (10) as well as a number of other genome annotation pipelines (11,12).

In the remainder of this article we describe developments in HAMAP since our last report in the Database Issue of Nucleic Acids Research. We also provide examples of how the careful manual curation of HAMAP profiles and associated rules can generate precise functional annotation for individual members of large and functionally diverse protein families.

ANNOTATION AND CONTENT

Refining HAMAP family profiles for increased specificity of functional annotation

HAMAP defines family membership of protein sequences using generalized profiles derived from manually curated multiple sequence alignments (MSAs) of trusted members (4,13). Precise functional annotation requires the careful definition of isofunctional protein families and functionally important residues—excluding other functional categories and closely related families curated in UniProtKB/Swiss-Prot. During curation of the multiple sequence alignment erroneous sequences and misaligned positions are corrected where necessary (described in (4), complete workflow ftp://ftp.expasy.org/databases/hamap/SOP_HAMAP_profile_creation.pdf included as supplementary file S1). Profiles are generated using the pftools package (available at http://web.expasy.org/pftools/) as described in (14,15). The specificity of the resulting profile may be modulated through the use of different pseudocounts, which assign scores to amino acid residues that have not been observed in the sequence alignments used to construct the profile (16). The values of these scores are derived from the PAM (Point Accepted Mutation) (17) and

```
ID   ENOPH_9FUNG             Unreviewed;       242 AA.
AC   3CFJ1;
DE   RecName: Full=Enolase-phosphatase E1;
DE            EC=3.1.3.77;
DE   AltName: Full=2,3-diketo-5-methylthio-1-phosphopentane phosphatase;

...

FT   METAL       197    197       Magnesium.
FT   BINDING     170    170       Substrate.
**
**   #################    INTERNAL SECTION    ##################
**HA Submitted Name: G8BDN2 - Hypothetical protein CPAR2_210240 (Candida parapsil...
**HA FAM; Method MF_03117; ENOPH; Trusted match; 63.485 (+7).
**HA FAM; Method MF_01681; MTNC; Weak match; 30.956 (-3.3).
**HA SAM; Annotated by HAMAP 1.10.1; MF_03117.3; MF_03117; 21-AUG-2014 12:08:13.
SQ   SEQUENCE   242 AA;  26760 MW;  171D6DDF4F2197F3 CRC64;
     MTIDTIILDI EGTVCPISFV KDTLFPYFIA KLPAELAKFD YPLNQVGSDN PIINILDNLP
     SNITESAQSV YDYLKNLVDS DIKDPVLKSL QGLIWTQGYD SGELKAPIYP DSIEFIESFP
     NRKRKIFIYS SGSIGAQKLL FGHVDNGTDK SIDLNKKLSG YFDITTAGHK TAYTSYSKIL
     KEIGKEDDPQ SVLFLSDNVD EVKAALKAGI ESYIVIRPGN APISDDDKSK YKTITAMSQL
     GI
//
```

Figure 3. Partial output of a HAMAP-Scan showing the additional information provided next to the actual annotations. The sequence of *Candida parapsilosis* hypothetical protein CPAR2_210240 (CCE43379.1) was submitted in FASTA format to HAMAP-Scan. The internal section in the output file contains information such as the submitted FASTA header, a trusted match (including the match score and the score difference to the trusted cut-off score) to profile MF_03117 (ENOPH), a weak match to profile MF_01681 (MTNC, the homologous bacterial family), as well as the information that the sequence has consequently been annotated by HAMAP rule MF_03117 associated with profile MF_03117. The full annotation produced for this sequence can be viewed in UniProtKB/TrEMBL record G8BDN2 for *C. parapsilosis* CPAR2_210240.

BLOSUM (BLOcks SUbstitution Matrix) (18) amino acid scoring matrices, which cover a wide range of evolutionary distances. Matrices tailored to shorter evolutionary distances will more strongly penalize substitutions that have not been observed, producing profiles that more faithfully reflect the observed diversity in the alignment—and which may better separate closely related subfamilies. There are of course limitations to this approach, and it is not always possible to generate HAMAP profiles that discriminate between very closely related sequences—one example, concerning certain subfamilies of sirtuins, is described below. The process of HAMAP family profile generation is iterative, and curators may modify the seed alignment, the profile construction parameters, and the threshold score for trusted family members until a profile with satisfactory specificity and sensitivity is achieved—based on the annotation of the matching UniProtKB/Swiss-Prot records. The parameters used for final profile generation are stored together with the seed alignment, so that profiles can be regenerated as needed.

HAMAP is continually updated, and HAMAP profiles and families may be modified, extended, or split as results from new phylogenetic analyses and experimental characterization data become available. A case in point is provided by the sirtuin family of proteins, whose members were thought to act exclusively as protein deacetylases (19,20). Phylogenetic analyses (using methods described in 21–25) suggest five families of sirtuins—classes I, II, III, IV and U (17) (see Figure 1). Class III sirtuins, including the human SIR5 protein (UniProtKB/Swiss-Prot record Q9NXA8), were recently found to exhibit both protein demalonylase and protein desuccinylase activity (26,27). The class III sirtuin of *Escherichia coli* (CobB, P75960) also functions as a protein desuccinylase (28), while that of *Plasmodium falciparum* (Sir2A, Q8IE47) hydrolyses medium and long chain fatty acyl groups from lysine residues (29), suggesting an ancient divergence of function in evolution. Specificity for these relatively bulky substrates may be conferred by a larger hydrophobic pocket and substrate-binding residues (Tyr-102 and Arg-105 in human SIR5) common to all class III sirtuins from all kingdoms of life (20,30). As part of the normal HAMAP workflow, all characterized sirtuin protein records in UniProtKB/Swiss-Prot were first updated (31). The existing HAMAP family profile for bacterial sirtuins (profile MF_01121) was modified to specifically match only the class III sirtuins, and new family profiles were created for classes II and U (profiles MF_01967 and MF_01968 respectively). HAMAP annotation rules for class III sirtuins were created that allow specific annotation of protein function and sequence features for both prokaryotic and eukaryotic sequences (rules MF_01121 and MF_03160 respectively). Class I and IV subfamilies are not currently treated by HAMAP, as these are further divided into subclasses (Ia, Ib, Ic and IVa, IVb, respectively), where each subclass contains multiple paralogs per species. Such complex duplications may be better addressed using methods that explicitly consider evolutionary history in the form of a phylogenetic tree. Other resources such as Pfam provide broad coverage of sirtuin family proteins (with a single signature PF02146) while a more restricted PIRSF signature (PIRSF037938) currently covers only the sirtuin subclass Ib members.

Table 1. The PFK family of proteins in HAMAP

HAMAP Profile AC	PFK family	Rule Scope	Annotations Substrate	Enzyme regulation	Subunit structure	Characterized template entries Species (UniProtKB AC), Kingdom, Substrate
ATP-dependent group I						
MF_00339	Prokaryotic clade B1	Bacteria	ATP	Allosteric activator = ADP and other NDPs, Allosteric inhibitor = PEP	Homotetramer	*E. coli* (P0A796), B, ATP *G. stearothermophilus* (P00512), B, ATP *L. lactis* (Q07636), B, ATP *Th. maritima* (Q9WY52), B, ATP *Th. thermophilus* (P21777), B, ATP
MF_03184	Eukaryotic two domain clade E	Eukaryota	ATP	Allosteric activator = ADP, AMP, fructose 2,6-bisphosphate, Allosteric inhibitor = ATP and citrate	Heterooctamer of 4 alpha and 4 beta chains (yeast), Homo- and heterotetramers (vertebrates) Homotetramer (others)	*S. cerevisiae* (P16861), E, ATP *S. cerevisiae* (P16862), E, ATP *D. discoideum* (P90521), E, ATP *H. sapiens* (P17858), E, ATP *H. sapiens* (P08237), E, ATP
PPi-dependent group II						
MF_01977	Clade P	Bacteria, Eukaryota	PPi	Non-allosteric	Homodimer or homotetramer	*P. freudenreichii* (P29495), B, PPi *M. alcaliphilum* (G4STG9), B, PPi *M. methanica* (Q3KSV5), B, PPi *M. balamuthi* (Q9NGP6), E, PPi
MF_01978	Prokaryotic clade B2	Bacteria	PPi	Non-allosteric	Homodimer	*M. capsulatus* (Q609I3), B, PPi *R. rubrum* (Q2RNU4), B, PPi *X. campestris* (B0RP51), B, PPi
MF_01979	Clade Short	Bacteria, Eukaryota	PPi	Non-allosteric	Homodimer (bacteria), Homotetramer (eukaryotes)	*Th. maritima* (Q9WYC5), B, PPi *N. fowleri* (Q27705), E, PPi *T. vaginalis* (A2E9H3), E, PPi
MF_01980	Clade Long	Bacteria, Eukaryota	PPi (or ATP?)	Non-allosteric, Allosteric activator = fructose 2,6-bisphosphate (plants)	Homodimer (bacteria) Tetramer of 2 alpha (regulatory) and 2 beta (catalytic) chains (plants), Homodimer or monomer (other eukaryotes)	*B. burgdorferi* (P70826), B, PPi *T. pallidum* (O83553), B, PPi *E. histolytica* (C4LZC2), E, PPi *A. thaliana* (Q8W4M5), E, PPi *A. thaliana* (Q9SYP2), E, Reg
MF_01981	Atypical ATP-dependent clade X	Bacteria, Eukaryota	ATP	Allosteric activator = AMP (eukaryota)	Homodimer (bacteria) Homotetramer (eukaryotes)	*A. methanolica* (Q8VU09), B, ATP *T. brucei* (O15648), E, ATP *E. histolytica* (Q27651), E, ATP *L. donovani* (Q9BIC6), E, ATP
MF_01976	Mixed-substrate group III	Archaea, Bacteria	PPi or ATP	Non-allosteric	Homodimer or homotetramer	*A. methanolica* (Q59126), B, PPi *S. coelicolor* (Q9L1L8), B, ATP

The 8 HAMAP profiles used to classify PFKs can be accessed at the HAMAP website by inserting the correct identifier into a URL of the form http://hamap.expasy.org/profile/<Profile AC> (e.g. http://hamap.expasy.org/profile/MF_00339 for prokaryotic clade B1 PFKs). The table summarizes the characteristics of the different subfamilies and the annotations that are propagated to matching target protein sequences. Characterized template proteins for each protein family are listed together with their origin (A = Archaea, B = Bacteria, E = Eukaryota) and the experimentally determined phosphoryl donor (ATP, PPi = inorganic phosphate, Reg = non-catalytic regulatory subunit). The full name and taxonomy of the species and the references describing protein characterization can be obtained from corresponding entries on the UniProt website via http://www.uniprot.org/uniprot/<UniProtKB AC> (e.g. http://www.uniprot.org/uniprot/P0A796 for *Escherichia coli* ATP-PFK pfkA).

HAMAP allows specific functional annotation within homologous protein families

The rule syntax used by HAMAP (described in http://hamap.expasy.org/unirule/unirule.html) allows for control statements that specify conditions–such as the occurrence of specific residues or motifs–for the application of annotation. These control statements provide a flexible means of fine-tuning the annotation of individual members of protein families, illustrated here using the 6-phosphofructokinase (PFK) family. PFK is a key regulatory enzyme of glycolysis that is present in all three domains of life. Despite this high level of conservation the enzyme has a remarkable evolutionary history, featuring a high rate of horizontal gene transfer and substitution in its active site (32). These substitutions have a profound impact on enzyme function; PFK family members with a glycine (G) at the active site catalyze the phosphorylation of D-fructose 6-phosphate to fructose 1,6-bisphosphate using adenosine triphosphate (ATP) (in the first committed step of glycolysis), while those with aspartate (D) use inorganic phosphate (PPi) as the phosphoryl donor in a reversible reaction that occurs in both glycolysis and gluconeogenesis (32–34).

HAMAP defines 8 PFK families in line with the currently accepted classification of PFKs (32,35) (Table 1). Several of the eight HAMAP families include both PPi-dependent and ATP-dependent members, suggesting that phosphoryl-donor specificity may have changed at multiple times during the evolution of the PFK superfamily. Figure 2 illustrates how this functional variation within families is treated by HAMAP using annotation rule MF_01976, which describes members of the mixed substrate PFK group III subfamily. The precise annotation that is applied to members of this family depends on the nature of the active site residue (D104 in the experimentally characterized template of *Amycolatopsis methanolica*—UniProtKB/Swiss-Prot record Q59126). Case statements within the rule specify the correct protein name, catalytic activity (including EC number), function, keywords, GO terms and other annotations for family members bearing either D or G at their active site. Sequences having neither of these residues are annotated as generic 6-phosphofructokinases of unknown

substrate-specificity. The example of PFK illustrates how a single residue may determine substrate specificity and enzyme function, but HAMAP rule syntax also allows conditional annotation based on the combination of multiple residues or sequence motifs. The methylthioadenosine (MTA) phosphorylases are one example, where conserved amino acid substitutions in the substrate binding pocket convert the substrate specificity of this enzyme from 6-aminopurine (EC 2.4.2.28) to 6-oxopurine nucleosides (EC 2.4.2.44 and EC 2.4.2.1) (described in MF_01963).

HAMAP statistics

Since our last publication in the NAR Database Issue 2013, we have added 203 new family profiles and 278 new annotation rules to HAMAP. As of 3 September 2014, HAMAP contains 1983 family classification profiles and 1998 annotation rules (a single HAMAP family profile may be associated with multiple HAMAP annotation rules, where each rule applies to a distinct taxonomic group). Through the UniRule pipeline, HAMAP provides annotations for 10,874,356 UniProtKB/TrEMBL sequence records (release 2014_08), which is around 13% of all sequence records in UniProtKB/TrEMBL, and 16% of the sequence records of each prokaryotic complete proteome. HAMAP provides 48% of all annotations and 90% of all sequence-specific feature annotations for the UniRule automatic annotation pipeline of UniProt. One of the strengths of HAMAP lies in the granularity and the comprehensiveness of its annotations, with each HAMAP rule providing over 16 annotations per UniProtKB/TrEMBL record on average.

WEBSITE

Improvements to the web-interface for HAMAP-Scan

Protein sequences can be classified and annotated using HAMAP through our HAMAP-Scan web service (http://hamap.expasy.org/hamap_scan.html). We provide a single-page, 3-step, dynamic submission form where required fields are clearly marked, and every field is accompanied by a short explanatory text. Each user choice dynamically updates the submission form, such that only necessary fields are displayed. The form allows submission of user sequences (FASTA) and UniProt sequence record identifiers or sequence accessions; users may submit individual sequences or whole proteome sequences. All submitted sequences are returned to the user in UniProtKB format in the order of submission, while protein sequences that have a trusted match to a HAMAP family profile are also annotated by the associated HAMAP rule. All result entries (including entries that are not annotated) contain an additional section with information on matches to HAMAP family profiles, including the profile accession number and identifier, the match quality (trusted or weak), and the match score (with the score difference to the trusted cut-off score of the profile in parenthesis) (Figure 3). HAMAP profiles are also available through InterProScan (36) provided by the InterPro Consortium (37), of which HAMAP is a member.

Accelerated HAMAP-Scan with pfsearchV3

To facilitate the use of HAMAP-Scan for the classification and annotation of large datasets such as whole proteome sequences we have implemented the improved version of the PROSITE search tool pfsearchV3 (38) for HAMAP. pfsearchv3 uses modern CPU instructions to exploit the capabilities of multicore processors and a new heuristic filter to rapidly score and select possible candidate matches, achieving speeds up to two orders of magnitude faster than the previous version of this algorithm. We plan to make the heuristic score thresholds for HAMAP profiles available to our users in the near future.

CONCLUSION

HAMAP provides accurate and detailed functional annotation for the exponentially growing population of uncharacterized protein sequences in public databases such as UniProtKB/TrEMBL, as well as tools and services for external users. HAMAP profiles allow the definition of isofunctional protein families of whatever size and scope according to current knowledge. HAMAP annotation rules provide fine-grained annotations for family members, based on the presence of specific functional residues (as illustrated here for the PFK families). The creation of family profiles and annotation rules in HAMAP is a manual effort performed by expert curators. Manual curation of the experimental literature in UniProtKB/Swiss-Prot is highly accurate (6), with expert curation of HAMAP profiles and rules specifically designed to avoid over-annotation through the careful definition of isofunctional protein families and functionally important residues. HAMAP annotations can be accessed via UniProtKB, or generated by users for their own protein or proteome sequences via the HAMAP-Scan service on the HAMAP website.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Anne Morgat and Marco Pagni for insightful comments and discussions on the scope and direction of HAMAP. We also thank Brigitte Boeckmann for critical reading of the manuscript and for help with the phylogenetic analysis of the sirtuin protein family.

FUNDING

Swiss Federal Government through the State Secretariat for Education, Research and Innovation; National Institutes of Health [U41HG006104]; Swiss National Science Foundation [JRP09 and JRP13]. Funding for open access charge: Swiss Federal Government through the State Secretariat for Education, Research and Innovation.
Conflict of interest statement. None declared.

REFERENCES

1. Gerlt,J.A., Allen,K.N., Almo,S.C., Armstrong,R.N., Babbitt,P.C., Cronan,J.E., Dunaway-Mariano,D., Imker,H.J., Jacobson,M.P.,

Minor,W. *et al.* (2011) Enzyme Function Initiative. *Biochemistry*, **50**, 9950–9962.
2. Anton,B.P., Chang,Y.C., Brown,P., Choi,H.P., Faller,L.L., Guleria,J., Hu,Z., Klitgord,N., Levy-Moonshine,A., Maksad,A. *et al.* (2013) The COMBREX project: design, methodology, and initial results. *PLoS Biol.*, **11**, e1001638.
3. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
4. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res.*, **41**, D584–D589.
5. Blake,J.A., Dolan,M., Drabkin,H., Hill,D.P., Li,N., Sitnikov,D., Bridges,S., Burgess,S., Buza,T., McCarthy,F. *et al.* (2013) Gene Ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
6. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
7. Bell,M.J., Collison,M. and Lord,P. (2013) Can inferred provenance and its visualisation be used to detect erroneous annotation? A case study using UniProtKB. *PLoS One*, **8**, e75541.
8. Gilks,W.R., Audit,B., De Angelis,D., Tsoka,S. and Ouzounis,C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.
9. UniProt Consortium. (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, doi:10.1093/nar/gku989.
10. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D.S., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
11. Darrasse,A., Carrere,S., Barbe,V., Boureau,T., Arrieta-Ortiz,M.L., Bonneau,S., Briand,M., Brin,C., Cociancich,S., Durand,K. *et al.* (2013) Genome sequence of Xanthomonas fuscans subsp. fuscans strain 4834-R reveals that flagellar motility is not a general feature of xanthomonads. *BMC Genomics*, **14**, 761.
12. Oakeson,K.F., Gil,R., Clayton,A.L., Dunn,D.M., von Niederhausern,A.C., Hamil,C., Aoyagi,A., Duval,B., Baca,A., Silva,F.J. *et al.* (2014) Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol. Evol.*, **6**, 76–93.
13. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
14. Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.*, **20**, 3–23.
15. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.*, **3**, 265–274.
16. Luthy,R., Xenarios,I. and Bucher,P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.*, **3**, 139–146.
17. Dayhoff,M.O., Schwartz,R. and Orcutt,B.C. (1978), *Atlas of protein sequence and structure*, Vol. **5**, pp. 345–358.
18. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–10919.
19. Sauve,A.A., Wolberger,C., Schramm,V.L. and Boeke,J.D. (2006) The biochemistry of sirtuins. *Annu. Rev. Biochem.*, **75**, 435–465.
20. North,B.J. and Verdin,E. (2004) Sirtuins: Sir2-related NAD-dependent protein deacetylases. *Genome Biol.*, **5**, 224.
21. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
22. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
23. Darriba,D., Taboada,G.L., Doallo,R. and Posada,D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164–1165.
24. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
25. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
26. Peng,C., Lu,Z., Xie,Z., Cheng,Z., Chen,Y., Tan,M., Luo,H., Zhang,Y., He,W., Yang,K. *et al.* (2011) The first identification of lysine malonylation substrates and its regulatory enzyme. *Mol. Cell. Proteomics*, **10**, M111 012658.
27. Du,J., Zhou,Y., Su,X., Yu,J.J., Khan,S., Jiang,H., Kim,J., Woo,J., Kim,J.H., Choi,B.H. *et al.* (2011) Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science*, **334**, 806–809.
28. Colak,G., Xie,Z., Zhu,A.Y., Dai,L., Lu,Z., Zhang,Y., Wan,X., Chen,Y., Cha,Y.H., Lin,H. *et al.* (2013) Identification of lysine succinylation substrates and the succinylation regulatory enzyme CobB in Escherichia coli. *Mol. Cell. Proteomics*, **12**, 3509–3520.
29. Zhu,A.Y., Zhou,Y., Khan,S., Deitsch,K.W., Hao,Q. and Lin,H. (2011) Plasmodium falciparum Sir2A preferentially hydrolyzes medium and long chain fatty acyl lysine. *ACS Chem. Biol.*, **7**, 155–159.
30. Frye,R.A. (2000) Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins. *Biochem. Biophys. Res. Commun.*, **273**, 793–798.
31. Poux,S., Magrane,M., Arighi,C.N., Bridge,A., O'Donovan,C. and Laiho,K. (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database (Oxford)*, bau016.
32. Bapteste,E., Moreira,D. and Philippe,H. (2003) Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. *Gene*, **318**, 185–191.
33. Moore,S.A., Ronimus,R.S., Roberson,R.S. and Morgan,H.W. (2002) The structure of a pyrophosphate-dependent phosphofructokinase from the Lyme disease spirochete Borrelia burgdorferi. *Structure*, **10**, 659–671.
34. Chi,A. and Kemp,R.G. (2000) The primordial high energy compound: ATP or inorganic pyrophosphate? *J. Biol. Chem.*, **275**, 35677–35679.
35. Muller,M., Lee,J.A., Gordon,P., Gaasterland,T. and Sensen,C.W. (2001) Presence of prokaryotic and eukaryotic species in all subgroups of the PP(i)-dependent group II phosphofructokinase protein family. *J. Bacteriol.*, **183**, 6714–6716.
36. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
37. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
38. Schuepbach,T., Pagni,M., Bridge,A., Bougueleret,L., Xenarios,I. and Cerutti,L. (2013) pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. *Bioinformatics*, **29**, 1215–1217.